Generation Z and the AI misinformation paradox: understanding a new digital vulnerability

Cecilia Cooley[1], Dr. Elizabeth Sperber[2]
[1]Student Contributor, University of Denver
[2]Faculty Advisor, Department of Political Science, University of Denver

Abstract

This study examines how and why Generation Z is particularly vulnerable to AI-generated content, and the implications of this vulnerability for democratic participation and information resilience. Drawing on empirical studies, survey data, and meta-analyses published between 2019 and 2025, this analysis synthesizes research on Gen Z's exposure to and interaction with AI-produced content across social media platforms. While this cohort has grown up in an increasingly digital age, multiple studies show that technological exposure and fluency do not reliably translate into digital literacy or resistance to misinformation. This vulnerability stems from three intersecting factors: high content volume, developmental traits favoring surface-level engagement, and algorithmic amplification of emotionally salient material, which collectively heighten susceptibility. Despite these risks, interventions such as prebunking and targeted media literacy training demonstrate measurable success in improving discernment and reducing belief in AI-generated falsehoods. These findings suggest that Gen Z's challenge is not a lack of access or awareness, but overexposure to AI-mediated information ecosystems that exploit attention and trust. Mitigating this vulnerability is essential for democratic resilience and the development of evidence-based policies.

Keywords: Artificial Intelligence (AI), misinformation, disinformation, Generation Z, prebunking, media literacy, cognitive bias, policy, education, digital resilience

1 INTRODUCTION

AI increasingly shapes the information youth consume, often in ways that undermine accuracy and critical evaluation. From deepfakes to algorithmically generated social media posts (content automatically produced or personalized by AI algorithms), these technologies circulate widely[1,2]. This proliferation poses novel challenges for Gen Z and the political systems in which they participate[2,3]. In democracies, citizens and their representatives rely on accurate information to self-govern effectively.

Born between 1997 and 2012, members of Gen Z grew up in a digitally saturated world. They engaged with AI tools and content earlier, more frequently, and more intuitively than any previous generation[1]. Although technological fluency is high, multiple studies show that early exposure to AI-mediated information environments does not guarantee resistance to mis/disinformation[2,4,5]. While some longitudinal evidence hints at potential life-course effects, this analysis primarily synthesizes current adolescent findings.

To situate Gen Z's experience, older cohorts (Millennials, born 1981–1996; Generation X, born 1965–1980) typically accessed digital media later in life and may rely on traditional news sources more heavily. This contrast underscores why early, immersive exposure to AI content can uniquely shape adolescent vulnerability[6].

Given these vulnerabilities, recent scholarship emphasizes educational and policy strategies that reduce the negative impacts of AI-generated mis/disinformation, particularly for younger audiences. Prebunking strategies, especially when delivered through interactive or game-based formats, have been shown to reduce susceptibility more effectively than disclosure labels, which demonstrate limited or inconsistent effects on belief and persuasion[7,8,9,10,11].

1.1 Key Terminology

Misinformation and disinformation encompass a broad range of content, from conspiracy theories and political deception to well-intentioned but inaccurate claims[2]. Misinformation refers to incorrect or misleading infor-

mation shared without intent to deceive, whereas disinformation is deliberately false content created to mislead, often spread covertly to influence public opinion or obscure the truth[12]. Freelon and Wells further refine this distinction by identifying three critical criteria for disinformation: deception, potential for harm, and intent to harm[3]. Recognizing these differences is essential for evaluating AI-generated content, which can simultaneously produce content that is misinformation and disinformation at a large scale, making it challenging for users to discern.

Artificial Intelligence (AI) is defined as computer systems capable of performing complex tasks typically requiring human reasoning, decision-making, or creativity[13]. AI includes technologies relevant to the spread of mis/disinformation, such as large language models (LLMs) that can recognize and generate text[14], voice cloning that reproduces a person's voice using recorded datasets[?], and AI-generated images and videos. While these tools are not inherently deceptive, their ability to produce content at scale and low cost makes them powerful vehicles for both misinformation and disinformation, particularly among youth[15,16]. Rapid generation and amplification of AI-produced content increases exposure to misleading information while also enabling the creation of content deliberately designed to deceive[2,17].

As noted above, Gen Z, the focus of this analysis, consists of individuals born between 1997 and 2012. Within this cohort, teens (ages 13–17) are particularly relevant for understanding the effects of AI generated content, as they are highly active online and still developing critical cognitive skills.

2 METHODS

This study employed a targeted literature review of peer-reviewed scholarship, policy reports, and reputable news sources examining youth, AI-generated content, and digital misinformation. The review specifically sought studies that explored patterns of social media use among Generation Z, cognitive vulnerabilities affecting susceptibility to misinformation, and interventions designed to mitigate AI-driven mis/disinformation, including key studies such as Roozenbeek and Van der Linden[7], Kyrychenko et al.[4], and Spearing et al.[8]. Keywords used in searches included: *"Generation Z"*, *"youth digital media"*, *"AI-generated content"*, *"misinformation"*, *"disinformation"*, *"prebunking"*, *"media literacy"*, and *"social media algorithms"*.

The initial search focused on peer-reviewed articles from databases such as Pew Research Center, JSTOR, Google Scholar, and Scopus, emphasizing research published between 2020 and 2025 to reflect recent developments in AI technologies. Key studies informing this review include investigations into social media habits and political information exposure[18,19], cognitive susceptibility to AI-generated misinformation[4,5], and experimental studies on prebunking and media literacy interventions[7,15], however, rather than examining these domains in isolation as prior research often does, this review synthesizes them to demonstrate how patterns of digital exposure, cognitive vulnerability, and intervention effectiveness interact to shape Generation Z's susceptibility to AI-driven misinformation.

Since U.S.-based programs are limited, European Union initiatives, including the Digital Services Act and AI Act, were also examined to provide a comparative perspective. Given the evolving nature of this field, conclusions remain tentative, particularly for culturally distinct subpopulations of Gen Z interacting with rapidly changing AI technologies.

2.1 Landscape and Exposure

Gen Z's pervasive use of social media and constant online connectivity amplifies their exposure to digital misinformation, creating unique vulnerabilities despite their digital nativity. According to a 202 Pew Research survey of 1,391 U.S. teens, nearly half reported being online "almost constantly," a 24% increase from a decade ago, while 96% reported daily internet use. TikTok, Instagram, and Snapchat remain the most popular platforms among U.S. teens, whereas Facebook and X use have declined sharply[18]. Similarly, a Pew Research Center survey of 1,453 U.S. teens ages 13–17 found that 95% have access to a smartphone, desktop, or laptop (Pew Research Center). Notably, 77% of U.S. young adults aged 18–34 reported using social media platforms such as X/Twitter, Instagram, TikTok, Reddit, YouTube, and Twitch as their primary sources of political information during the 2024 election cycle[19]. Together, these findings demonstrate that today's youth are both highly connected (frequent access) and heavily reliant (primary source for information and social interaction) on digital media.

Algorithmic curation further shapes Gen Z's online environment. Platforms tend to prioritize sensational, emotionally charged, or highly engaging content, increasing the visibility of misleading or polarizing material[1,2]. Influencers—individuals with substantial online followings— also significantly shape information consumption. When political or social commentary is embedded within entertainment or lifestyle content, teens may struggle to distinguish factual reporting from opinion or sponsored messaging[2]. These patterns create conditions in which misinformation can spread rapidly and broadly, particularly among youth with underdeveloped critical evaluation skills.

Table 1 Seven Types of Mis/Disinformation, Reproduced information from Claire Wardle, *"Fake news. It's complicated"*, First Draft, 2017.

Type	Definition	Harm Level	Category
Satire or Parody	No intention to cause harm but has potential to fool	Least Harmful	Misinformation
False Connection	When headlines, visuals, or captions don't support the content	Low	Misinformation
Misleading Content	Misleading use of information to frame an issue or individual	Moderate	Misinformation
Fake Context	When genuine content is shared with false contextual information.	Moderate	Disinformation
Imposter Content	When genuine sources are impersonated	High	Disinformation
Manipulated Content	When genuine information or imagery is manipulated to deceive	High	Disinformation
Fabricated Content	New content that is 100% false, made to deceive and do harm	Most Harmful	Disinformation

A 2025 Misinformation Susceptibility Test (MIST) asked participants to classify twenty headlines (ten real, ten false) as "fake" or "real." Gen Z participants demonstrated the lowest accuracy across age groups, misclassifying a significant proportion of false headlines[4]. Despite growing up immersed in technology, many teens lack formal training in evaluating sources, including structured media literacy, critical news analysis, or instruction in recognizing manipulated or synthetic content[5]. While some individuals may have encountered such training in higher education or professional settings, it is not consistently embedded in K–12 curricula, leaving teens particularly vulnerable. The MIST also revealed that Gen Z expressed low confidence in identifying misinformation, aligning self-perception with actual performance[4]. This combination of high exposure and low confidence creates a cycle in which teens remain easily influenced by misleading content.

Experimental evidence highlights the real-world consequences of these vulnerabilities. In one study, 52% of youth participants interpreted a grainy video filmed in Russia as evidence of ballot tampering in the 2016 U.S. primaries, even though it lacked verification or context[2]. The video depicted indistinct footage of individuals handling ballots and moving around a polling location, which participants misinterpreted as fraudulent activity. Across participants, 97% failed to fact-check sources or consider potential biases, and 66% could not differentiate news stories from sponsored content[2]. Globally, similar patterns persist: a UNICEF survey of ten countries found that up to three-quarters of youth reported feeling unable to judge the veracity of online information[2].

AI-generated misinformation adds an additional layer of concern. Teens (defined here as individuals aged 13–17) encounter AI-written captions, synthetic images, and automated videos at scale, often crafted to mimic the style and tone of credible sources, making discernment especially difficult[1,2]. AI can also promote cognitive offloading, a measurable process in which users rely on surface-level cues such as headlines, visuals, or perceived source credibility rather than engaging in deeper analytical reasoning[1]. These adolescents rarely verify sources when encountering AI content, reflecting both limited critical evaluation skills and the overwhelming volume of AI-generated material[1].

These findings underscore the need for targeted educational and policy interventions, including age-appropriate media literacy curricula, explicit instruction on identifying AI-generated content, platform transparency requirements, and safeguards limiting algorithmic amplification of misleading information. Without such measures, Gen Z's high exposure to AI-driven content may continue to outpace their ability to critically evaluate it.

2.2 Cognitive Vulnerability Mechanisms

While artificial intelligence offers notable benefits across sectors like health care, education, and finance, its rapid integration into youth information ecosystems amplifies cognitive vulnerabilities and susceptibility to misinformation[16]. Moreover, the sheer volume of AI-produced content can overwhelm cognitive capacity, making it increasingly difficult for teens to distinguish credible material from misinformation.

AI systems rely on machine learning algorithms to detect patterns in large datasets rather than being explicitly programmed. This autonomy has made content production faster, cheaper, and less dependent on human input[16]. Major news organizations, such as The Washington Post, Reuters, and Bloomberg, and tech platforms including TikTok and YouTube, have integrated generative AI tools for headline generation, automated

summaries, or image and video enhancement[1,16]. However, because these systems learn from human-created data, they often replicate existing perceptual, emotional, and social biases. For example, if AI content is trained on politically polarized posts, it may disproportionately present emotionally charged political content, normalizing extreme perspectives[16].

Adolescents' cognitive development amplifies these vulnerabilities. Executive function, impulse control, and metacognitive monitoring, which are essential for evaluating credibility, detecting bias, and reflecting on information, are still developing in teens[16,20]. Consequently, repeated exposure to biased AI content can shape implicit attitudes, reinforce confirmation bias, and normalize distorted perspectives. If an AI system repeatedly presents hyper partisan headlines as credible news, this may lead adolescents to perceive such framing as typical or authoritative, even when it is misleading.

While users bring preexisting biases to online environments, AI can encode and amplify these biases through engagement-optimized ranking, probabilistic content generation, and training on biased datasets[1,16]. Biased or unverified content combined with passive scrolling, where users minimally engage or reflect, creates ideal conditions for misinformation to flourish[1]. Social media algorithms, optimized to capture attention and promote shareable material, routinely amplify sensational or emotionally charged content, including clickbait, conspiratorial rhetoric, and deceptive narratives[2].

AI intensifies these vulnerabilities by enabling the rapid production of hyper-realistic images and videos that mimic journalistic aesthetics, dramatically increasing both the scale and perceived credibility of misleading visual content[1]. AI-generated news-style videos may replicate visual markers of credibility, such as professional formatting, authoritative narration, and branded graphics, making synthetic content difficult to distinguish from legitimate reporting. Visual information is processed far more rapidly than written communication, and youth are less likely to employ critical reasoning when encountering videos or images[2]. This effect is particularly pronounced for AI-generated visual content, which can include deepfakes or hyper-realistic media that mimic credible sources, bypassing critical scrutiny. Consequently, even well-intentioned adolescents may unknowingly internalize misleading narratives, reinforcing stereotypes or misconceptions over time[16,20].

The convergence of low production costs, high accessibility, and limited factchecking creates a high-risk online environment for young users. For adolescents with developing brains, repeated exposure to manipulative or false narratives can reshape worldview, influence political attitudes, and erode trust in institutions, threatening democratic participation itself[20]. For instance, a teen overwhelmed by AI-generated stories vilifying a social group or swayed by a deepfake political video may withdraw from civic engagement altogether. This dynamic illustrates the duality of AI, highlighting how it can both educate and inspire, for example by providing accessible explanations of complex topics, generating personalized learning content, and supporting interactive engagement with news and civic information, while also subtly distorting beliefs, particularly among impressionable youth. Such examples are not merely hypothetical but are already observable in today's media landscape[20]. Understanding mechanisms such as algorithmic amplification, AI-generated content that mimics credible sources, and users' reliance on surface-level cues rather than critical evaluation is essential for designing interventions that both leverage AI's educational potential and mitigate its risks.

2.3 Intervention Mechanisms

Addressing youth vulnerability to AI-generated misinformation requires targeted interventions. As existing research highlights both the potential and limitations of strategies such as disclosure labels, prebunking, and media literacy programs, one consistent conclusion across modern scholarship on AI generated mis/disinformation is the urgent need for reform. Policymakers face a difficult balance: They must protect youth from harmful content while respecting the right to freedom of expression and access to information. The latter "can be infringed by over-zealous attempts, including regulations, to restrict access to online content and communities"[2]. To date, three main intervention strategies have been studied: disclosure labels, prebunking, and media literacy programs. While these approaches are often discussed separately, they should be understood as complementary rather than mutually exclusive. As summarized in Table 2, research suggests that each approach has both strengths and limitations. Collectively, these approaches aim to improve digital discernment, strengthen critical reasoning, and reduce the likelihood that adolescents will unknowingly share or internalize false information.

Disclosure labels are defined as "explicit messages, icons, or text that accompany media and alert viewers that the content was created or modified by an artificial intelligence system"[21]. Although designed to increase transparency, these labels are often ignored or distrusted by teens. Gallegos and coauthors found that while 94.6% of youth participants in a study recognized authorship labels, the labels had no significant effect on message accuracy judgments, attitude change, or sharing intentions[9]. Similarly, Schmalzle and coauthors found through a second study that labeling messages as AI-generated influenced evaluation but not ranking, with a slight bias against AI-created messages[21]. Li and

Table 2 Overview of Major Intervention Types.

Intervention	Definition	Effectiveness	Strengths	Limitations
Disclosure Labels	Labels or icons indicating content was AI-generated	Low–moderate	Increase transparency and awareness	Often ignored; minimal behavioral impact; may reduce credibility of true content
Prebunking (Inoculation Theory)	Exposure to weakened misinformation techniques to build resistance	High (short-term)	Builds cognitive resilience; effective across ages; works well when interactive or game-based	Effects fade without reinforcement; requires facilitation
Media Literacy Programs	Education on analyzing, evaluating, and creating digital content responsibly	Moderate–high (long-term)	Promotes critical thinking; scalable in schools; adaptable for youth	Resource-intensive; uneven implementation; requires curriculum updates

Yang likewise concluded that labeling had no meaningful effect on perceived accuracy, message credibility, or sharing intention, though labels did not negatively impact overall platform trust[10]. These studies suggest that while labeling may increase awareness, it is insufficient to alter cognitive or behavioral responses on its own, particularly among adolescents who tend to be highly visually oriented and exposed to continuous content streams[2].

These weak effects should not be interpreted as a reason to abandon disclosure labels entirely, but rather as a signal that the labels may require refinement, through either clearer design, integration with interactive cues, or reinforcement through education to increase effectiveness.

A related study of U.S. and U.K. youth revealed that labeling headlines as AI-generated reduced perceived accuracy and willingness to share them, even when the headlines were true or human-written[11]. However, this "AI aversion" effect was three times weaker than labeling content as false, suggesting that audiences assume AI-generated headlines are unsupervised by humans[11]. These results underscore that labeling should be implemented carefully to avoid stigmatizing legitimate AI content. Collectively, current research calls for further study on how disclosure effects vary by content type, region, and demographic factors before they can be considered a reliable mitigation tool.

Prebunking, or inoculation theory, involves warning individuals about potential misinformation and exposing them to a weakened form of misleading techniques to build cognitive resistance[7]. Evidence on prebunking's effectiveness among youth is encouraging, though its effects tend to fade without reinforcement. Spearing and coauthors found through a study that source-focused inoculation reduced overall trust in AI-generated information but did not significantly diminish the influence of specific misleading articles[8]. In another study, however, combining inoculation with debunking eliminated misinformation effects entirely[7]. Fulsher and coauthors found that using generative AI itself for prebunking through model fine-tuning, prompt design, and classroom integration, showed strong potential but required human facilitation to ensure accuracy and ethical oversight[15]. This highlights the dual role of AI as both a source of misinformation and a potential tool for cultivating critical reasoning. As Jia and coauthors argue, all education stakeholders "must be prepared to thoughtfully navigate GenAI's dual potential as both a source of and solution to misinformation"[17].

Youth-oriented prebunking studies further highlight interactive learning. Dangol and coauthors developed *AI Puzzlers*, a system that helps children identify reasoning errors in generative AI. Even younger children (ages six and under) who were not fluent readers could detect inconsistencies in AI generated solutions, leading to reflective discussions about "how AI thinks"[22]. Likewise, Roozenbeek and Van der Linden's "fake news game" trained players to use six misinformation tactics, polarization, emotion, conspiracy, trolling, deflection, and impersonation, and found improved resistance to misinformation across age, ideology, and education level[7]. Gamification and interactive methods engage multiple cognitive pathways, encouraging adolescents to recognize manipulation strategies actively rather than passively consuming content. Collectively, this evidence suggests that prebunking, especially when gamified and reinforced over time, is one of the most effective tools to strengthen youth misinformation resilience.

Media literacy programs represent a third major intervention. These structured educational frameworks teach students how to access, analyze, evaluate, and create media responsibly[23]. Results are mixed but promising. Yim and Su highlight the growth of AI literacy initiatives using age appropriate tools such as Google's

Teachable Machine, *Learning ML*, and *Machine Learning for Kids*, which enhance both engagement and soft-skills acquisition[24]. Jia and coauthors proposed a holistic AI literacy framework consisting of three dimensions: AI awareness, AI mechanics, and AI impacts, each encompassing skills from understanding AI applications to responsible practice[17]. Li and coauthors found that while teachers generally understand AI's core functions, they face challenges accommodating students' varying AI skill levels; encouragingly, opportunities for AI literacy learning did not significantly vary by socioeconomic status[10]. These programs aim to build durable skills for critical evaluation, fostering long-term resilience against emerging AI-generated misinformation.

All three approaches – disclosure labels, prebunking, and media literacy programs – show partial effectiveness. Disclosure labels enhance transparency but yield minimal behavioral change and risk being misunderstood or ignored. Prebunking is the most empirically supported, showing strong short-term effects, especially through gamified learning, though it depends on reinforcement and human facilitation. Media literacy programs provide comprehensive frameworks and long-term potential but face scalability and resource barriers. Taken together, embedding prebunking games within school curricula and sustaining media literacy education may create the most robust defense against AI-driven mis/disinformation among adolescents.

3 POLICY CONTEXT

Current policy frameworks addressing AI-generated misinformation vary across regions. The European Union has implemented strategies such as the Digital Services Act and the AI Act, aiming to increase platform accountability, enhance transparency of algorithmic content, and protect vulnerable populations, including youth[1]. Together, these policies provide a model for balancing innovation with consumer protection[2]. However, transferring EU approaches to the U.S. context is complicated by institutional differences. The U.S. places a stronger emphasis on First Amendment protections, which operates under decentralized regulatory structures, and often relies on industry self regulation, which may limit direct applicability of EU frameworks[1,2]. Adapting EU inspired frameworks to the U.S. would therefore require careful consideration of legal and political constraints to ensure both effectiveness and respect for freedom of expression.

4 CONCLUSIONS

As AI technologies evolve at an unprecedented pace, effective policy and educational interventions must simultaneously advance to protect Gen Z from escalating risks of AI-driven misinformation. As demonstrated throughout this analysis, scholars widely agree on two central points: (1) Gen Z is highly vulnerable to AI-generated misinformation and disinformation due to factors such as content volume, sophistication, developmental stage, accessibility, and media habits; and (2) concrete policy and intervention efforts to mitigate these effects and close generational literacy gaps remain insufficient. Promising interventions, particularly prebunking strategies delivered through interactive, gamified, or educational formats, show measurable short-term benefits, but scalability and long-term durability remain limited, highlighting the need for coordinated institutional support. Early policy efforts, particularly those establishing accountability for AI developers and protections for children, mark important progress but require sustained expansion and enforcement. Given that Gen Z represents the future electorate and workforce, policymakers and local institutions must act urgently to strengthen these initiatives.

In the near term, schools represent the most direct and scalable setting for equipping youth with the critical tools to navigate digital environments. As AI becomes an embedded feature of daily life, educators must treat media and AI literacy as core competencies on par with traditional subjects. Embedding interactive prebunking exercises, critical evaluation tasks, and scenario-based learning into curricula can reinforce resilience, allowing students to apply analytical reasoning in real-world contexts. Integrating prebunking and media literacy through interactive, experiential learning can help students cultivate discernment, skepticism, and resilience from an early age. Over time, embedding such practices into formal education may lay the foundation for a more critically literate society.

Meanwhile, it is critical for policymakers to accelerate legislation addressing AI-enabled misinformation, which has a disproportionate impact on youth. Since regulatory processes often lag technological development, action must be proactive rather than reactive. Comprehensive AI governance anchored in transparency, safety, and accountability should remain a top priority to protect Gen Z and global democratic integrity from the destabilizing effects of AI-generated misinformation. In practice, this requires policies that are adaptive, regularly reviewed, and informed by empirical evidence on youth engagement with AI content. Continuous collaboration between lawmakers, educators, researchers, and platform developers is essential to translate policy into effective, real-world safeguards.

4.1 Future Directions

Despite growing evidence of Gen Z's susceptibility to AI-generated misinformation, significant gaps remain, highlighting the urgent need for longitudinal, cross-cultural, and intervention-focused research. Specifically,

future research should prioritize prolonged studies to assess how Gen Z's misinformation exposure and resilience evolve over the life course and as AI systems become more sophisticated. Tracking these trajectories can help identify critical periods for intervention and inform the design of sustainable educational strategies. More extensive cross-cultural analyses are also needed to develop a deeper understanding of how differences in education systems, media environments, and policy frameworks affect susceptibility across nations.

Another critical avenue involves evaluating the sustainability of prebunking and media literacy programs. To what extent, if at all, do game-based and interactive interventions maintain their benefits over longer periods of time, and for whom? Research should also examine how hybrid approaches combining digital and in-person interaction, can reinforce learning outcomes and support adaptive coping strategies against evolving AI misinformation techniques. Additionally, scholars will likely continue to explore how AI can be leveraged as a defensive tool to combat the negative effects of AI-generated mis/disinformation.

Finally, policymakers and developers must collaborate to design transparent, youth-centered AI governance frameworks that protect against manipulation while empowering young users as informed digital citizens. Such frameworks should incorporate principles of accountability, inclusivity, and evidence-based oversight, ensuring that youth are not only protected but also equipped to engage critically with digital media. Advancing these efforts will be key to closing the generational vulnerability gap and ensuring that future information ecosystems promote resilience, critical awareness, and democratic integrity.

5 ACKNOWLEDGEMENTS

The author gratefully acknowledges the University of Denver's College of Arts, Humanities, and Social Sciences for enabling this research. I would also like to thank Professor Elizabeth Sperber for her guidance and insightful feedback throughout this project.

6 EDITOR'S NOTES

This article was peer-reviewed.

REFERENCES

[1] *How's Life for Children in the Digital Age?* (OECD Publishing, 2025).

[2] Howard, P. N., Neudert, L.-M., Prakash, N. & Vosloo, S. Digital misinformation/disinformation and children. *UNICEF Office of Global Insight and Policy* (2021).

[3] Freelon, D. & Wells, C. Disinformation as political communication. *Political Communication* **37**, 145–156 (2020).

[4] Kyrychenko, Y. *et al.* Profiling misinformation susceptibility. *Personality and Individual Differences* **241**, 113177 (2025).

[5] Dolan, E. Gen z and conservatives show higher misinformation susceptibility, large-scale study finds. *Social Psychology* (2025).

[6] Sidoti, O., Park, E. & Gottfried, J. About a quarter of u.s. teens have used chatgpt for schoolwork – double the share in 2023. *Pew Research Center* (2025).

[7] Roozenbeek, J. & van der Linden, S. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* **5**, 65 (2019).

[8] Spearing, E. R. *et al.* Countering ai-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science* **12** (2025).

[9] Gallegos, I. O. *et al.* Labeling messages as ai-generated does not reduce their persuasive effects (2025).

[10] Li, F. & Yang, Y. Impact of artificial intelligence–generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based, randomized, controlled experiment. *JMIR Formative Research* **8**, e60024 (2024).

[11] Altay, S. & Gilardi, F. People are skeptical of headlines labeled as ai-generated, even if true or human-made, because they assume full ai automation. *PNAS Nexus* **3** (2024).

[12] Association, A. P. Misinformation and disinformation .

[13] NASA.gov. What is artificial intelligence? .

[14] Amazon. What is llm (large language model)? .

[15] Fulsher, A., Pagkratidou, M. & Kendeou, P. Genai and misinformation in education: a systematic scoping review of opportunities and challenges. *AI SOCIETY* (2025).

[16] Glickman, M. & Sharot, T. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour* **9**, 345–359 (2024).

[17] Jia, K., Leung, T. H. M., Cheung, N. Y. I., Li, Y. & Yu, J. Developing a holistic ai literacy framework for children. *ACM Transactions on Computing Education* (2025).

[18] Faverio, M. & Sidoti, O. Teens, social media and technology 2024 (2024).

[19] Siegel-Stechler, K., Hilton, K. & Medina, A. Youth rely on digital platforms, need media literacy to access political information. *Center for Information Research on Civic Learning and Engagement* (2025).

[20] Rohman, D. F. Y. *et al.* The influence of artificial intelligence on information integrity: A media literacy approach for young people. *International Journal of Environmental Sciences* **11**, 1022–1034 (2025).

[21] Lim, S. & Schmälzle, R. The effect of source disclosure on evaluation of ai-generated messages. *Computers in Human Behavior: Artificial Humans* **2**, 100058 (2024).

[22] Dangol, A. *et al.* "ai just keeps guessing": Using arc puzzles to help children identify reasoning errors in generative ai. In *Proceedings of the 24th Interaction Design and Children*, 444–464 (ACM, 2025).

[23] Potter. Media literacy. *Sage Publishing* (2019).

[24] Yim, I. H. Y. & Su, J. Artificial intelligence (ai) learning tools in k-12 education: A scoping review. *Journal of Computers in Education* (2024).